Electronic medical records and genomics (eMERGE) network exploration in cataract: Several new potential susceptibility loci

Molecular Vision 2014; 20:1281-1295 <http://www.molvis.org/molvis/v20/1281>
Received 15 March 2012 | Accepted 17 September 2014 | Published 19 September 2014

Electronic medical records and genomics (eMERGE) network exploration in cataract: Several new potential susceptibility loci

Marylyn D. Ritchie,[1] Shefali S. Verma,[1] Molly A. Hall,[1] Robert J. Goodloe,[2] Richard L. Berg,[3] Dave S. Carrell,[4] Christopher S. Carlson,[5] Lin Chen,[6] David R. Crosslin,[7,8] Joshua C. Denny,[9,10] Gail Jarvik,[7,11] Rongling Li,[12] James G. Linneman,[13] Jyoti Pathak,[14] Peggy Peissig,[13] Luke V. Rasmussen,[15] Andrea H. Ramirez,[10] Xiaoming Wang,[9] Russell A. Wilke,[9,16] Wendy A. Wolf,[17] Eric S. Torstenson,[2] Stephen D. Turner,[18] Catherine A. McCarty[19]

[1]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA; [2]Center for Human Genetics Research, Vanderbilt University, Nashville, TN; [3]Biomedical Informatics Research Center, Biostatistics, Marshfield Clinic Research Foundation, Marshfield, WI; [4]Group Health Research Institute, Seattle, WA; [5]Fred Hutchinson Cancer Research Center, Seattle, WA; [6]Ophthalmology, Marshfield Clinic Research Foundation, Marshfield, WI; [7]Division of Medical Genetics, University of Washington, Seattle, WA; [8]Department of Biostatistics, University of Washington, Seattle, WA; [9]Departments of Biomedical Informatics, Vanderbilt University, Nashville, TN; [10]Department of Medicine, Vanderbilt University, Nashville, TN; [11]Departments of Medicine and Genome Sciences, University of Washington, Seattle, WA; [12]Office of Population Genomics, National Human Genome Research Institute, Bethesda, MD; [13]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI; [14]Department of Biomedical Informatics, Mayo Clinic College of Medicine, Rochester, MN; [15]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University, Chicago, IL; [16]IMAGENETICS at Sanford Medical Center, Fargo, ND and Department of Internal Medicine, University of North Dakota, Fargo, ND; [17]Division of Genetics and Genomics, Boston Children's Hospital and Department of Pediatrics, Harvard Medical School, Boston, MA; [18]Public Health Sciences, University of Virginia, Charlottesville, VA; [19]Essentia Institute of Rural Health, Duluth, MN

Purpose: Cataract is the leading cause of blindness in the world, and in the United States accounts for approximately 60% of Medicare costs related to vision. The purpose of this study was to identify genetic markers for age-related cataract through a genome-wide association study (GWAS).

Methods: In the electronic medical records and genomics (eMERGE) network, we ran an electronic phenotyping algorithm on individuals in each of five sites with electronic medical records linked to DNA biobanks. We performed a GWAS using 530,101 SNPs from the Illumina 660W-Quad in a total of 7,397 individuals (5,503 cases and 1,894 controls). We also performed an age-at-diagnosis case-only analysis.

Results: We identified several statistically significant associations with age-related cataract (45 SNPs) as well as age at diagnosis (44 SNPs). The 45 SNPs associated with cataract at $p<1\times10^{-5}$ are in several interesting genes, including *ALDOB*, *MAP3K1*, and *MEF2C*. All have potential biologic relationships with cataracts.

Conclusions: This is the first genome-wide association study of age-related cataract, and several regions of interest have been identified. The eMERGE network has pioneered the exploration of genomic associations in biobanks linked to electronic health records, and this study is another example of the utility of such resources. Explorations of age-related cataract including validation and replication of the association results identified herein are needed in future studies.

Cataract is the leading cause of blindness in the world [1,2], is the leading cause of vision loss in the United States [3], and accounts for approximately 60% of Medicare costs related to vision [4]. Summary prevalence estimates indicate that 17.2% of Americans aged 40 years and older have cataract in either eye and 5.1% have pseudophakia or aphakia (previous cataract surgery). In addition to the implications for healthcare delivery and healthcare costs, cataract has been shown to be associated with falls and increased mortality [5-12], possibly because of associated systemic conditions. Women have a slightly higher risk of having cataract than men [13]. With increased life expectancy, the number of cataract cases and cataract surgeries is expected to increase dramatically unless primary prevention strategies can be developed and successfully implemented.

Several genetic loci have also been linked to cataract as an independent phenotypic trait. An extensive body of literature has addressed the role of genetics in childhood cataract [14], and it has been hypothesized that these same genes may be plausible candidates for age-related cataract [15]. It has been suggested that as many as 40 genes may be involved in age-related cataract [16]. Evidence for a major gene has been identified for cortical [17] and nuclear [18,19]

Correspondence to: Marylyn Ritchie, Pennsylvania State University, Center for Systems Genomics, The Huck Institutes for the Life Sciences, Department of Biochemistry and Molecular Biology, 512 Wartik Laboratory, University Park, PA 16802; Phone: (814) 863-5107; FAX: (814) 863-6699; email: marylyn.ritchie@psu.edu

cataract, with heritability estimates of 58% [20] and 48% [21], respectively. A whole genome STR scan conducted in families in Wisconsin revealed a major locus for age-related cortical cataract on chromosome 6p12-q12 [22], and specific candidate genes that have been studied include *galactokinase* (Gene_ID: 2584; OMIM: 604313) [23,24], *apolipoprotein E* (Gene_ID: 348; OMIM: 107741) [25], *glutathione S-transferase* (Gene_ID: 2944; OMIM: 138350)[26], *N-acetyltransferase 2* (Gene_ID: 10; OMIM: 612182) [27,28], and estrogen metabolism genes [29]. Two recent studies found an association between the *EPHA2* gene (Gene_ID: 1969; OMIM: 176946) and cataract [30,31].

Higher body mass index (BMI) has been shown in many studies to increase risk of cortical and posterior subcapsular (PSC) cataract (odds ratio [OR] = 1.5–2.5) [32-38]. A recent study found that nuclear cataract was not associated with obesity but was associated with the *FTO* obesity gene (Gene_ID: 79068; OMIM: 610966) in an Asian population [39]. Although familial aggregation studies have shown a potential role for gene and environment interactions in nuclear cataract [40,41], research in this area is limited. The association of glutathione S-transferase with cataract has been shown to be modified by smoking [42] and sunlight exposure [43]. No whole genome association SNP studies of age-related cataract in unrelated individuals have been reported in the medical literature. The purpose of this study was to conduct a genome-wide association study (GWAS) for age-related cataract and to prioritize top hits for further follow-up.

METHODS

Phenotypic data: The National Human Genome Research Institute (NHGRI)-funded electronic medical records and genomics (eMERGE) network implemented an electronic phenotype algorithm to select cataract cases and controls [44]. Cataracts as a condition were selected by Marshfield Clinic as its primary eMERGE phenotype, and the algorithm, which uses diagnostic and procedure codes, was developed by the Marshfield Clinic Personalized Medicine Research Project (PMRP) investigators [45]. The five sites in eMERGE-I include Marshfield Clinic, Group Health Research Institute, Vanderbilt University, Mayo Clinic, and Northwestern University. This study included four of the sites: Marshfield Clinic, Group Health Research Institute, Vanderbilt University, and Mayo Clinic. Using an algorithm for a specific phenotype, each participating site extracted study samples for a specific disease or phenotype from the electronic health records (EHR). Once samples had been selected and genotyped, they were available for phenotyping with additional algorithms. Thus, the cataract algorithm was deployed across the network. The cases and the controls had to meet the following inclusion criteria: The cases were age 50 years and older at the time of diagnosis or surgery, and the controls were age 50 years or older at the time of the most recent eye exam and had had an eye exam within the previous 5 years. The controls had no diagnostic codes for cataract or evidence of cataract surgery. The cases were identified as "surgical" or "diagnosis only." Surgical cases had undergone a cataract extraction in at least one eye. The diagnosis-only cases were required to have either cataract diagnoses on two or more dates or have one diagnosis date and natural language processing and optical character recognition (NLP/OCR) find one or more inclusion cataract terms. Cataract type was extracted from the notes using natural language processing and optical character recognition with validation through manual chart abstraction [45,46].

Genotypic data: Genome-wide genotyping has been performed on approximately 17,000 samples across the network at the Broad Institute and at the Center for Inherited Disease Research (CIDR) using the Illumina 660W-Quad or 1M-Duo Beadchips (CIDR, Baltimore, MD). For this particular study, which includes predominantly individuals of European descent, we used only the Illumina 660W-Quad platform. This platform consists of 561,490 SNPs and 95,876 intensity-only probes. Genotyping calls were made at either CIDR or Broad using BeadStudio version 3.3.7. The eMERGE Cataract dataset pre-quality control (QC) included 7,535 DNA samples and 344 HapMap controls: 3,968 Marshfield Clinic, 2,379 Group Health, 986 Mayo, and 202 Vanderbilt BioVU. Data were cleaned using the eMERGE QC pipeline developed by the eMERGE Genomics Working Group [47]. This process includes evaluation of the sample and marker call rate, gender mismatch, duplicate and HapMap concordance, batch effects, Hardy–Weinberg equilibrium, sample relatedness, and population stratification. After QC, 530,101 SNPs and 7,397 samples were used for analysis (see Table 1 for distribution by site). All genotype data and a detailed QC report for each individual site, as well as the merged eMERGE dataset, can be found on dbGaP, and the detailed eMERGE QC pipeline can be found in [47,48].

Statistical analyses: Single-locus tests of association were performed using PLINK [49] assuming an additive genetic model for all 530,101 SNPs in a total of 7,397 unrelated individuals (5,503 cases and 1,894 controls). We calculated principal components using the EIGENSTRAT program [50] and thus adjusted our analyses for the first three principal components (PCs) to avoid any spurious associations that can be caused due to population stratification. EIGENSTRAT is based on principal components analysis and is used to detect

Table 1. Descriptive statistics on eMERGE cataract data set.

Study sample	Site			Number	Number Missing	Total
				7397	0	7397
	Marshfield Clinic		Total			3914 (52.91%)
			Cases	2557		
			Controls	1357		
	Mayo Clinic		Total			952 (12.87%)
			Cases	606		
			Controls	346		
	Group Health		Total			2346 (31.72%)
			Cases	2235		
			Controls	111		
	Vanderbilt		Total			185 (2.50%)
			Cases	105		
			Controls	80		
	White			7109 (96.11%)		
Race	Black			114 (1.54%)	0	7397
(p=0.1157)						
	Other			174 (2.35%)		
Case-Control Cataract			Cases	5503	0	7397
			Controls	1894		
Cataract Age at Diagnosis (Case only)		Mean±SD	70.50±8.09	7296	101	7397
		Median	71			
		IQR(25%,75%)	(66,76)			
		Range	35 - 136			
Sex	Cases		Male	2401	0	5503
(p=0.7684)			Female	3102		
	Controls		Male	819	0	1894
			Female	1075		
Birthdate Year**	Cases	Mean±SD	2.3211±1.01	5462	41	5503
(p<0.0001)		Median	2			
		IQR(25%,75%)	(1,3)			
		Range	1.000 - 5.000			
	Controls	Mean±SD	4.07±0.936	1893	1	1894

Study sample	Site			Number	Number Missing	Total
		Median	4			
		IQR(25%,75%)	(4,5)			
		Range	1.000 - 6.000			
Diabetes (p<0.0001)	Cases		Yes	921	2613	5503
			No	1969		
	Controls		Yes	306	323	1894
			No	1265		

** Birthdate Year denotes decade of birth where 1=1910, 2=1920, 3=1930, 4=1940, 5=1950, 6=1960

Molecular Vision 2014; 20:1281-1295 <http://www.molvis.org/molvis/v20/1281>

and correct for population stratification in genome-wide association studies. Thus, we present the results of the analysis adjusted by principal components 1–3 (PC1–3).

We also performed an age-at-diagnosis association analysis using cases only. Age at diagnosis is defined as the age when the first cataract diagnosis was made in the electronic health record. We performed unadjusted analysis and adjusted for PC1–3 using linear regression in PLINK. In Table 2 and Table 3, we report all p values $<1\times10^{-5}$. All associations identified by our analyses are suggestive and must be replicated in independent datasets because the signals did not reach a Bonferroni corrected genome-wide statistical significance level.

RESULTS

Figure 1 shows the Manhattan plots for the single locus tests of association for cataract case control adjusted (Figure 1A) and age-at-diagnosis adjusted (Figure 1B) and Figure 2 shows the corresponding QQ plots for each GWAS analysis. Our top hits in the adjusted case-control analysis include *gigaxonin* (*GAN;* Gene_ID: 8139, OMIM: 605379; p value = 2.42×10^{-6}), which encodes a member of the cytoskeletal Broad-Complex, Tramtrack, and Bric a brac (BTB/kelch) repeat family. The encoded protein plays a role in neurofilament architecture and is involved in mediating the ubiquitination and degradation of some proteins. Defects in this gene are a cause of giant axonal neuropathy (*GAN*). Other potential interesting findings include *DNER* (Gene_ID: 92737; OMIM: 607299; p value = 1.87×10^{-5}), which encodes for the Delta and Notch-like epidermal growth factor-related receptor, and *EHHADH* (Gene_ID: 1962; OMIM: 607037; p value = 2.80×10^{-5}) encodes for enoyl-CoA, hydratase/3-hydroxyacyl CoA dehydrogenase. Myocyte-specific enhancer factor 2C also known as MADS box transcription enhancer factor 2, polypeptide C is a protein that in humans is encoded by the *MEF2C* gene (Gene_ID: 4208; OMIM: 600662; p value = 7.26×10^{-5}). *MEF2C* upregulates the expression of the homeodomain transcription factors DLX5 and DLX6, two transcription factors that are necessary for craniofacial development [51]. This could be another interesting link to cataracts.

Several SNPs in or near *ALDOB* (Gene_ID: 229; OMIM: 612724; p value = 2.46×10^{-6}), which encodes for aldolase B, fructose-bisphosphate, were also associated with cataracts in our GWAS analysis. Mutations in this gene result in an autosomal recessive disorder of fructose intolerance, and cases of cataract have been reported in the first decade of life [52]. Another interesting associated gene is *MAP3K1* (Gene_ID: 4214; OMIM: 600982; p value = 1.33×10^{-5}), a functional mitogen-activated protein kinase kinase kinase 1. Molecular signatures of *MAP3K1* have been shown to be important in embryonic eyelid closure in the mouse [53]. In total, 45 SNPs were statistically significant at $p<10^{-5}$ or smaller.

In the age-at-diagnosis analysis, our top hits include *ACSS3* (Gene_ID: 79611; OMIM: 614356; p value = 6.39×10^{-7}), which is *acyl-CoA synthetase short-chain family member 3*; EPHA4 (p value = 7.03×10^{-5}), ephrin type-A receptor 4, which is a protein that in humans is encoded by the *EPHA4* gene (Gene_ID: 2043; OMIM: 602188). This gene belongs to the ephrin receptor subfamily of the protein-tyrosine kinase family, along with *EPHA2*. EPH and EPH-related receptors have been implicated in mediating developmental events, especially in the nervous system [54].

DISCUSSION

This study is the first genome-wide association study in age-related cataract reported in the literature. Cataract in type 2 diabetes has been investigated, and a region on chromosome 3p14.4–3p14.2 was identified in a Han Chinese population [55]. The five SNPs identified in that study do not show evidence of association in our eMERGE cataract GWAS. It is difficult to interpret these results, however, because age-related cataracts and cataracts in type 2 diabetics may be two different phenotypes, which may have disparate etiologies. In addition, our dataset does not have an overwhelming number of individuals with type 2 diabetes (see Table 1); thus, we were underpowered to explore this specific type of association. Other previously published research on gene mapping in cataracts supports a linkage region on chromosome 1 [56] and association with *EPHA2* [30,31]. In our GWAS, we did not see evidence for association with *EPHA2*, although we did see association with *EPHA4*. One significant difference in this study is the phenotyping of cases and controls based on electronic health records (EHR) in population-based cohorts, rather than family-based samples. However, our study in addition to the literature supports the suggestion of cataract-susceptibility loci on chromosome 1. Replication studies and larger sample sizes are needed to validate and confirm these findings.

Although the eMERGE network has demonstrated the utility of electronic phenotyping in EHR for several traits [57-61], there are inherent challenges with this approach. For ophthalmic conditions specifically, the abundance of EHR coded information is extremely limited or, in some health systems, absent. Thus, sophisticated phenotyping strategies must be established [45,46] Still, the success of the EHR and biobank approach for association studies is unprecedented. The ability to perform multiple GWAS simultaneously with no additional genotyping is an enormous benefit [58]. Once a

Table 2. PC adjusted case-control association analysis results.

CHR	SNP	Reference Allele	Case MAF	OR	P value	Gene	Left Gene	Right Gene	Type of Variant
16	rs8044853	T	0.335	0.7099	2.42E-06	NA	GAN	CMIP	NA
9	rs1929494	T	0.4391	1.217	2.46E-06	LOC100129210	ALDOB	C9orf125	intron
22	rs926937	A	0.045	0.8525	6.09E-06	NA	LOC100130624	MN1	NA
16	rs9927153	A	0.2391	0.8359	9.38E-06	NA	GAN	CMIP	NA
16	rs2098753	G	0.3183	0.8106	1.06E-05	NA	GAN	CMIP	NA
5	rs9292118	A	0.2659	1.193	1.17E-05	NA	LOC441073	MAP3K1	NA
1	rs16853148	A	0.059	1.28	0.000012	NA	PRDM2	RP1–21O18.1	NA
5	rs13178221	T	0.243	1.203	1.33E-05	NA	LOC441073	MAP3K1	NA
9	rs882809	T	0.3823	0.7482	1.48E-05	LOC100129210	ALDOB	LOC100129210	near-gene-5
10	rs9299674	G	0.3242	0.7436	1.53E-05	NA	LOC441550	LOC439953	NA
10	rs4301693	C	0.1521	1.184	1.84E-05	NA	LOC441550	LOC439953	NA
2	rs10197959	A	0.4305	0.8409	1.87E-05	DNER	PID1	LOC100130031	intron
16	rs1563655	A	0.3251	0.8514	2.04E-05	NA	GAN	CMIP	NA
2	rs4853633	T	0.1937	1.241	2.17E-05	NA	MSTN	MGC13057	NA
15	rs8027435	T	0.4498	1.235	2.23E-05	NA	ARRDC4	LOC728459	NA
3	rs13074058	C	0.0789	0.8423	0.000028	LOC285382	VPS8	EHHADH	intron
10	rs549676	C	0.4961	1.219	3.19E-05	NA	PITRM1	KLF6	NA
2	rs10864871	C	0.2922	0.7878	3.26E-05	NA	hCG_2045614	LOC728241	NA
6	rs9405313	A	0.1204	0.7855	3.31E-05	NA	LY86	RP11–320C15.1	NA
4	rs4695885	C	0.3323	1.222	3.96E-05	NA	LOC100128266	FBXO8	NA
4	rs2015977	A	0.4608	0.5185	4.08E-05	NA	LOC391656	LOC100131441	NA
16	rs310011	G	0.4267	0.8545	4.39E-05	NA	GAN	CMIP	NA
3	rs3732933	A	0.0718	1.181	4.42E-05	EHHADH	C3orf70	EIF2S2P2	reference
12	rs7963343	C	0.1752	1.203	4.49E-05	LOC100129881	CRADD	LOC441644	intron
20	rs6073358	T	0.0897	0.8249	4.57E-05	JPH2	TOX2	C20orf111	intron
18	rs7244678	C	0.0764	1.183	6.02E-05	IMPA2	MPPE1	LOC646044	intron
19	rs7252479	A	0.0516	0.8323	6.02E-05	ZNF578	LOC441862	ZNF808	intron
15	rs1993976	A	0.4469	0.7933	6.74E-05	NA	ARRDC4	LOC728459	NA
3	rs17008958	A	0.1439	0.7515	7.02E-05	EIF4E3	FOXP1	GPR27	intron
13	rs943386	G	0.324	1.258	7.13E-05	NA	LOC646208	LOC100130029	NA
17	rs4531770	C	0.1407	0.8437	0.000072	NA	hCG_1644301	FLJ37644	NA
5	rs3850653	A	0.2327	1.178	7.26E-05	NA	MEF2C	LOC729011	NA

CHR	SNP	Reference Allele	Case MAF	OR	P value	Gene	Left Gene	Right Gene	Type of Variant
1	rs10746432	A	0.4345	0.8413	7.53E-05	HHAT	LOC100129235	KCNH1	intron
5	rs160044	T	0.3105	1.232	7.62E-05	MEF2C	LOC645323	LOC729011	intron
3	rs1447899	T	0.2838	1.246	0.00008	EIF4E3	FOXP1	GPR27	intron
12	rs4831958	T	0.0711	0.8469	8.04E-05	NA	LOC100130336	LOC100131830	NA
4	rs6814129	G	0.4445	1.217	8.11E-05	NA	MRPS36P2	LOC644325	NA
9	rs12347205	A	0.3934	1.21	8.47E-05	NA	IL6RL1	OR7E31P	NA
9	rs951611	T	0.0095	0.8008	9.12E-05	NA	LOC286239	LOC401497	NA
1	rs4951508	T	0.2343	0.7486	9.58E-05	HHAT	LOC100129235	KCNH1	intron
6	rs9379053	A	0.1076	0.7166	9.65E-05	NA	LY86	RP11–320C15.1	NA
20	rs1337906	C	0.3422	1.179	9.84E-05	NA	RPL41P1	ST13P	NA
9	rs2148996	T	0.4654	0.8493	9.86E-05	NA	LOC392358	GAS1	NA
19	rs7247032	T	0.3922	0.7675	9.91E-05	NA	LOC100130084	USP29	NA
8	rs4268128	A	0.215	0.6761	9.97E-05	NA	TNFRSF10B	TNFRSF10C	NA

Molecular Vision 2014; 20:1281-1295 <http://www.molvis.org/molvis/v20/1281>

Table 3. PC adjusted age-at-diagnosis association analysis results.

CHR	SNP	Reference Allele	Case MAF	Beta	P value	Gene	Left Gene	Right Gene	Type of Variant
12	rs12296937	G	0.0267	−1.08	6.39E-07	ACSS3	LIN7A	PPFIA2	intron
12	rs2574730	A	0.0371	−1.003	3.04E-06	ACSS3	LIN7A	PPFIA2	intron
12	rs769056	T	0.0369	0.6454	3.39E-06	ACSS3	LIN7A	PPFIA2	intron
12	rs11835432	T	0.1937	0.6667	7.64E-06	NA	LOC100132564	LOC644489	NA
1	rs207145	T	0.1237	−0.6722	7.9E-06	NA	LOC645506	GOT2L1	NA
12	rs2593270	A	0.2593	−1.263	1.06E-05	NA	LOC100132564	LOC644489	NA
12	rs2656824	G	0.2529	0.7466	1.19E-05	NA	LOC100132564	LOC644489	NA
15	rs4965818	G	0.3444	−0.8272	1.37E-05	SNRPA1	SELS	PCSK6	intron
12	rs337656	T	0.2225	−0.7027	1.45E-05	NA	LOC643264	CLLU1OS	NA
12	rs10778791	G	0.0354	−0.7015	2.08E-05	ACSS3	LIN7A	PPFIA2	intron
2	rs12612521	C	0.2144	−0.6601	0.000024	NA	LOC728241	LOC100131284	NA
2	rs10932058	C	0.4981	−0.7464	2.68E-05	NA	LOC100132132	LOC100132669	NA
15	rs748696	G	0.4491	−0.6481	2.98E-05	KIAA1199	FAM108C1	LOC100128570	intron
15	rs1524876	T	0.4568	0.7457	3.57E-05	MTMR10	MTMR15	TRPM1	intron
15	rs4778856	G	0.4667	0.7429	3.77E-05	KIAA1199	FAM108C1	LOC100128570	intron
9	rs2229594	T	0.1722	0.7258	3.95E-05	BAAT	LOC347275	LOC100128665	utr-3
16	rs933717	T	0.435	0.6402	0.000041	FBXO31	LOC730018	MAP1LC3B	intron
1	rs6663771	G	0.4138	0.652	4.24E-05	NA	SPATA17	RRP15	NA
15	rs1432442	G	0.0913	0.6621	0.000043	MAP2K1	ATP5J2P6	SNAPC5	intron
5	rs2468475	T	0.473	−0.6838	4.59E-05	NA	LOC100128659	LOC729862	NA
4	rs2406040	G	0.266	0.8714	4.59E-05	NA	LOC646316	LOC729578	NA
4	rs2406041	C	0.2591	−1.213	0.000051	NA	LOC646316	LOC729578	NA
2	rs13414831	G	0.2974	−2.441	5.34E-05	NA	UBR3	MYO3B	NA
20	rs864184	A	0.2301	−1.864	0.000054	PHACTR3	LOC645605	SYPC2	intron
4	rs10517073	T	0.4173	−2.027	5.98E-05	ANAPC4	ZCCHC4	LOC645433	intron
18	rs578026	C	0.3204	−2.022	0.000061	CLUL1	CETN1	C18orf56	intron
2	rs16857804	G	0.2944	−0.8456	6.24E-05	NA	UBR3	MYO3B	NA
2	rs4560089	G	0.3525	1.146	6.35E-05	NA	LOC100130842	MRPL50P1	NA
12	rs934078	A	0.1049	−1.048	6.57E-05	NA	OSTF1P	TBX3	NA
1	rs1416156	A	0.4182	0.8037	6.96E-05	NA	SPATA17	RRP15	NA
2	rs617222	A	0.2478	0.8032	7.03E-05	NA	LOC100129746	EPHA4	NA
4	rs2897305	G	0.2646	0.8959	7.06E-05	NA	LOC646316	LOC729578	NA

CHR	SNP	Reference Allele	Case MAF	Beta	P value	Gene	Left Gene	Right Gene	Type of Variant
20	rs6070943	A	0.1765	−0.6294	7.23E-05	PHACTR3	LOC645605	SYCP2	intron
1	rs991007	T	0.1118	0.656	7.26E-05	INADL	TM2D1	L1TD1	intron
12	rs12099972	A	0.0821	1.132	7.71E-05	NA	LOC100129881	LOC441644	NA
2	rs9309489	A	0.2321	−0.6522	7.86E-05	NA	TACR1	FAM176A	NA
14	rs1742707	A	0.4406	0.6657	7.97E-05	NA	CPSF2	SLC24A4	NA
17	rs9908117	C	0.2586	−0.7235	8.27E-05	NA	LOC100128284	WSCD1	NA
9	rs7874443	C	0.3182	−0.6617	8.34E-05	NA	GOLM1	LOC100130433	NA
2	rs10195113	T	0.0657	−0.7182	8.41E-05	NA	SLC8A1	LOC729984	NA
5	rs1472606	G	0.3331	−0.69	0.000085	NA	SFXN1	HRH2	NA
18	rs7227421	G	0.0358	−1.679	9.02E-05	GNAL	LOC729602	CHMP1B	intron
10	rs4388822	T	0.0717	−0.7606	9.14E-05	NA	LOC439992	GRID1	NA
5	rs2277939	A	0.3473	−0.8382	9.34E-05	SAP30L	GALNT10	HAND1	intron

Molecular Vision 2014; 20:1281-1295 <http://www.molvis.org/molvis/v20/1281>

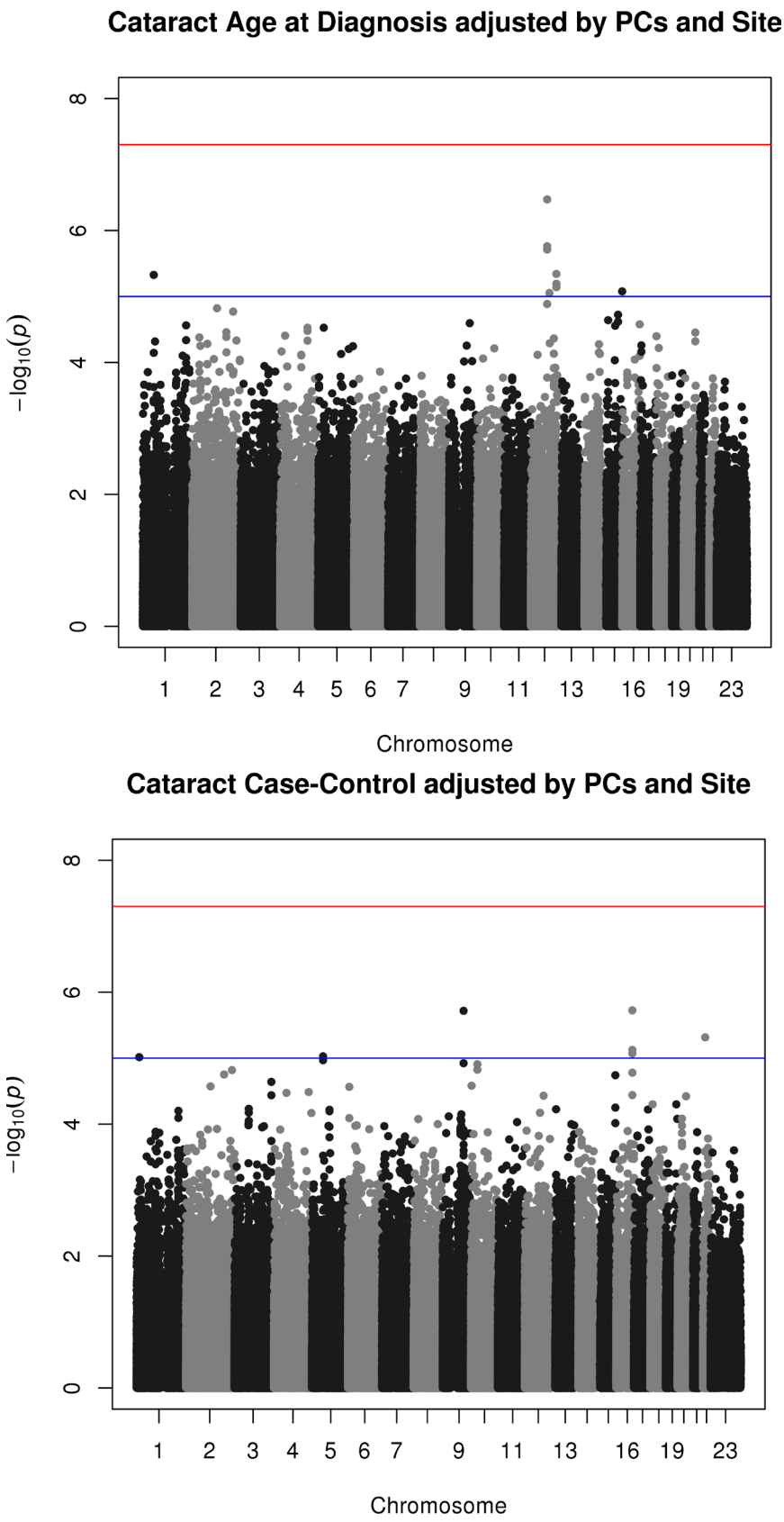
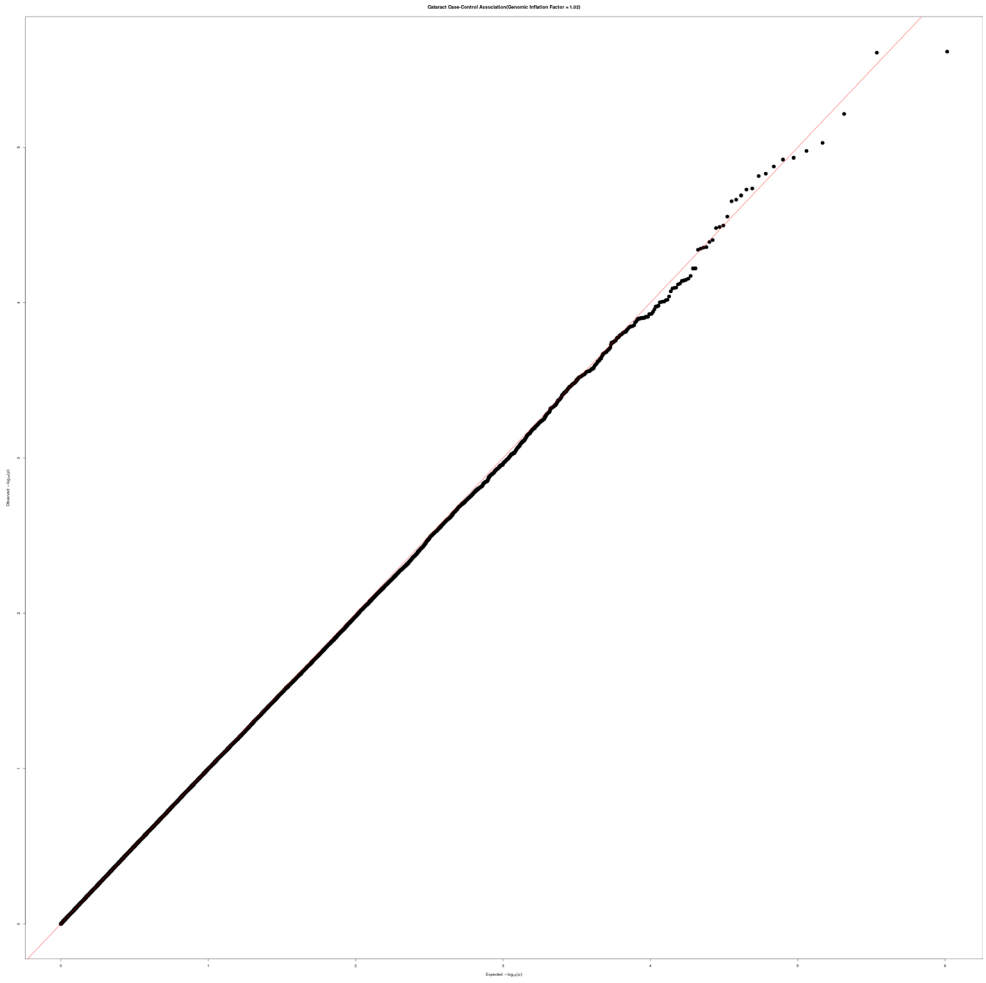


Figure 1. Genome-wide association study Manhattan plots for cataract and age-at-cataract-diagnosis. **A**: Case-control adjusted by first three principal components and site where eMERGE data was collected. **B**: Age-at-diagnosis adjusted by first three principal components and site where eMERGE data were collected.



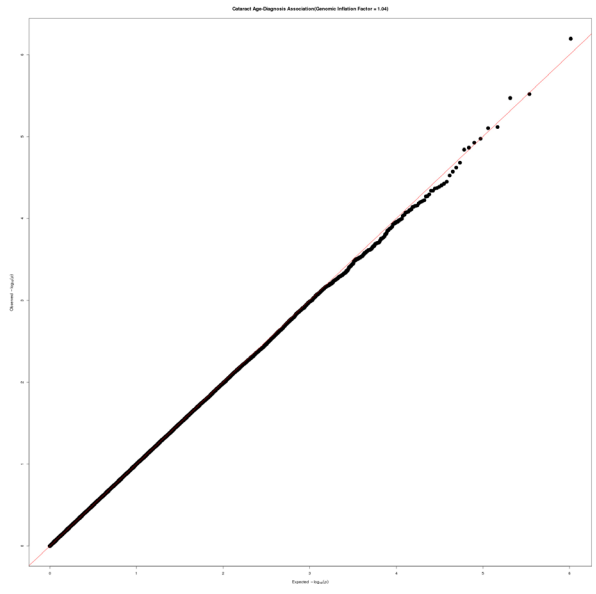


Figure 2. Quantile-quantile plot for analysis adjusted by the first three principal components and site where eMERGE data were collected.

set of patient samples has been genotyped on a genome-wide association platform, those data can be reused for multiple additional genotype-phenotype association studies. In particular, the eMERGE network has done quite a bit of this for quantitative traits and clinical laboratory variables such as cholesterol [60], red-blood cell indices [59], and white blood cell count [57]. The additional effort is expended on creating electronic phenotyping algorithms, rather than collecting samples and genotyping. Thus, this is an enormous resource for subsequent genotype-phenotype association studies.

Future explorations of age-related cataract include validating and replicating the association results identified herein. Unfortunately, because of the sample size and limited power by stratifying cases and controls by the eMERGE site, we did not have the opportunity to replicate these findings within eMERGE. The goal is to identify a similar study population where these results can be explored. In addition, we are beginning to investigate the role of gene–gene and gene–environment interactions associated with cataracts [62]. Due to the complexity of the trait, we hypothesize that the genetic architecture will be similar to that of other complex traits: multigenic with a combination of genetic and environmental interactions.

As demonstrated by this and other studies, the beauty of using an electronic health record is the ability to reuse genotyped samples for various phenotypes. The eMERGE network has clearly demonstrated the success of this study design, and continues to demonstrate the strengths and limitations of this approach.

ACKNOWLEDGMENTS

The eMERGE Network was initiated and funded by NHGRI, with additional funding from NIGMS through the following grants: U01HG004610 (Group Health Cooperative); U01HG004608 (Marshfield Clinic); U01HG04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01HG04603 (Vanderbilt University, also serving as the Coordinating Center); U01HG006389 (Essentia Institute of Rural Health). The Northwest Institute of Medical Genetics is also supported by a State of Washington Life Sciences Discovery Fund award.

REFERENCES

1. Thylefors B, Negrel A-D. Available data on blindness. Geneva, Switzerland: World Health Organization; 1994.
2. Black A, Wood J. Vision and falls. Clin Exp Optom 2005; 88:212-22. [PMID: 16083415].
3. Congdon N, O'Colmain B, Klaver CCW, Klein R, Muñoz B, Friedman DS, Kempen J, Taylor HR, Mitchell P. Causes and prevalence of visual impairment among adults in the United States. Arch Ophthalmol 2004; 122:477-85. [PMID: 15078664].
4. Ellwein LB, Urato CJ. Use of eye care and associated charges among the Medicare population: 1991–1998. Arch Ophthalmol 2002; 120:804-11. [PMID: 12049587].
5. Podgor MJ, Cassel GH, Kannel WB. Lens changes and survival in a population-based study. N Engl J Med 1985; 313:1438-44. [PMID: 4058547].
6. Minassian DC, Mehra V, Johnson GJ. Mortality and cataract: findings from a population-based longitudinal study. Bull World Health Organ 1992; 70:219-23. [PMID: 1600582].
7. West SK, Muñoz B, Istre J, Rubin GS, Friedman SM, Fried LP, Bandeen-Roche K, Schein OD. Mixed lens opacities and subsequent mortality. Arch Ophthalmol 2000; 118:393-7. [PMID: 10721963].
8. Wang JJ, Mitchell P, Simpson JM, Cumming RG, Smith W. Visual impairment, age-related cataract, and mortality. Arch Ophthalmol 2001; 119:1186-90. [PMID: 11483087].
9. Williams SL, Ferrigno L, Mora P, Rosmini F, Maraini G. Baseline cataract type and 10-year mortality in the Italian-American Case-Control Study of age-related cataract. Am J Epidemiol 2002; 156:127-31. [PMID: 12117703].
10. Reidy A, Minassian DC, Desai P, Vafidis G, Joseph J, Farrow S, Connolly A. Increased mortality in women with cataract: a population based follow up of the North London Eye Study. Br J Ophthalmol 2002; 86:424-8. [PMID: 11914212].
11. Clemons TE, Kurinij N, Sperduto RD. Associations of mortality with ocular disorders and an intervention of high-dose antioxidants and zinc in the Age-Related Eye Disease Study: AREDS Report No. 13. Arch Ophthalmol 2004; 122:716-26. [PMID: 15136320].
12. Knudtson MD, Klein BEK, Klein R. Age-related eye disease, visual impairment, and survival: the Beaver Dam Eye Study. Arch Ophthalmol 2006; 124:243-9. [PMID: 16476894].
13. Congdon N, Vingerling JR, Klein BEK, West S, Friedman DS, Kempen J, O'Colmain B, Wu S-Y, Taylor HR. Prevalence of cataract and pseudophakia/aphakia among adults in the United States. Arch Ophthalmol 2004; 122:487-94. [PMID: 15078665].
14. Reddy MA, Francis PJ, Berry V, Bhattacharya SS, Moore AT. Molecular genetic basis of inherited cataract and associated phenotypes. Surv Ophthalmol 2004; 49:300-15. [PMID: 15110667].
15. Moore AT. Understanding the molecular genetics of congenital cataract may have wider implications for age related cataract. Br J Ophthalmol 2004; 88:2-3. [PMID: 14693758].
16. Hejtmancik JF, Kantorow M. Molecular genetics of age-related cataract. Exp Eye Res 2004; 79:3-9. [PMID: 15183095].
17. Heiba IM, Elston RC, Klein BE, Klein R. Evidence for a major gene for cortical cataract. Invest Ophthalmol Vis Sci 1995; 36:227-35. [PMID: 7822150].

18. Heiba IM, Elston RC, Klein BE, Klein R. Genetic etiology of nuclear cataract: evidence for a major gene. Am J Med Genet 1993; 47:1208-14. [PMID: 8291558].

19. . The Framingham Offspring Eye Study Group. Familial aggregation of lens opacities: the Framingham Eye Study and the Framingham Offspring Eye Study. Am J Epidemiol 1994; 140:555-64. [PMID: 8067349].

20. Hammond CJ, Duncan DD, Snieder H, De Lange M, West SK, Spector TD, Gilbert CE. The heritability of age-related cortical cataract: the twin eye study. Invest Ophthalmol Vis Sci 2001; 42:601-5. [PMID: 11222516].

21. Hammond CJ, Snieder H, Spector TD, Gilbert CE. Genetic and environmental factors in age-related nuclear cataracts in monozygotic and dizygotic twins. N Engl J Med 2000; 342:1786-90. [PMID: 10853001].

22. Iyengar SK, Klein BEK, Klein R, Jun G, Schick JH, Millard C, Liptak R, Russo K, Lee KE, Elston RC. Identification of a major locus for age-related cortical cataract on chromosome 6p12-q12 in the Beaver Dam Eye Study. Proc Natl Acad Sci USA 2004; 101:14485-90. [PMID: 15452352].

23. Okano Y, Asada M, Fujimoto A, Ohtake A, Murayama K, Hsiao KJ, Choeh K, Yang Y, Cao Q, Reichardt JK, Niihira S, Imamura T, Yamano T. A genetic factor for age-related cataract: identification and characterization of a novel galactokinase variant, "Osaka," in Asians. Am J Hum Genet 2001; 68:1036-42. [PMID: 11231902].

24. Maraini G, Hejtmancik JF, Shiels A, Mackay DS, Aldigeri R, Jiao XD, Williams SL, Sperduto RD, Reed G. Galactokinase gene mutations and age-related cataract. Lack of association in an Italian population. Mol Vis 2003; 9:397-400. [PMID: 12942049].

25. Zetterberg M, Zetterberg H, Palmér M, Rymo L, Blennow K, Tasa G, Juronen E, Veromann S, Teesalu P, Karlsson J-O, Höglund K. Apolipoprotein E polymorphism in patients with cataract. Br J Ophthalmol 2004; 88:716-8. [PMID: 15090431].

26. Juronen E, Tasa G, Veromann S, Parts L, Tiidla A, Pulges R, Panov A, Soovere L, Koka K, Mikelsaar AV. Polymorphic glutathione S-transferases as genetic risk factors for senile cortical cataract in Estonians. Invest Ophthalmol Vis Sci 2000; 41:2262-7. [PMID: 10892871].

27. Tamer L, Yilmaz A, Yildirim H, Ayaz L, Ates NA, Karakas S, Oz O, Yildirim O, Atik U. N-acetyltransferase 2 phenotype may be associated with susceptibility to age-related cataract. Curr Eye Res 2005; 30:835-9. [PMID: 16251120].

28. Meyer D, Parkin DP, Seifart HI, Maritz JS, Engelbrecht AH, Werely CJ, Van Helden PD. NAT2 slow acetylator function as a risk indicator for age-related cataract formation. Pharmacogenetics 2003; 13:285-9. [PMID: 12724621].

29. Lee S-M, Tseng L-M, Li A-F, Liu H-C, Liu T-Y, Chi C-W. Polymorphism of estrogen metabolism genes and cataract. Med Hypotheses 2004; 63:494-7. [PMID: 15288375].

30. Shiels A, Bennett TM, Knopf HLS, Maraini G, Li A, Jiao X, Hejtmancik JF. The EPHA2 gene is associated with cataracts linked to chromosome 1p. Mol Vis 2008; 14:2042-55. [PMID: 19005574].

31. Jun G, Guo H, Klein BEK, Klein R, Wang JJ, Mitchell P, Miao H, Lee KE, Joshi T, Buck M, Chugha P, Bardenstein D, Klein AP, Bailey-Wilson JE, Gong X, Spector TD, Andrew T, Hammond CJ, Elston RC, Iyengar SK, Wang B. EPHA2 is associated with age-related cortical cataract in mice and humans. PLoS Genet 2009; 5:e1000584-[PMID: 19649315].

32. Glynn RJ, Christen WG, Manson JE, Bernheimer J, Hennekens CH. Body mass index. An independent predictor of cataract. Arch Ophthalmol 1995; 113:1131-7. [PMID: 7661746].

33. Hiller R, Podgor MJ, Sperduto RD, Nowroozi L, Wilson PW, D'Agostino RB, Colton T. A longitudinal study of body mass index and lens opacities. The Framingham Studies. Ophthalmology 1998; 105:1244-50. [PMID: 9663229].

34. Caulfield LE, West SK, Barrón Y, Cid-Ruzafa J. Anthropometric status and cataract: the Salisbury Eye Evaluation project. Am J Clin Nutr 1999; 69:237-42. [PMID: 9989686].

35. Schaumberg DA, Glynn RJ, Christen WG, Hankinson SE, Hennekens CH. Relations of body fat distribution and height with cataract in men. Am J Clin Nutr 2000; 72:1495-502. [PMID: 11101477].

36. Weintraub JM, Willett WC, Rosner B, Colditz GA, Seddon JM, Hankinson SE. A prospective study of the relationship between body mass index and cataract extraction among US women and men. Int J Obes Relat Metab Disord 2002; 26:1588-95. [PMID: 12461675].

37. Jacques PF, Moeller SM, Hankinson SE, Chylack LT Jr, Rogers G, Tung W, Wolfe JK, Willett WC, Taylor A. Weight status, abdominal adiposity, diabetes, and early age-related lens opacities. Am J Clin Nutr 2003; 78:400-5. [PMID: 12936921].

38. Kuang T-M, Tsai S-Y, Hsu W-M, Cheng C-Y, Liu J-H, Chou P. Body mass index and age-related cataract: the Shihpai Eye Study. Arch Ophthalmol 2005; 123:1109-14. [PMID: 16087846].

39. Lim LS, Tai E-S, Aung T, Tay WT, Saw SM, Seielstad M, Wong TY. Relation of age-related cataract with obesity and obesity genes in an Asian population. Am J Epidemiol 2009; 169:1267-74. [PMID: 19329528].

40. Congdon N, Broman KW, Lai H, Munoz B, Bowie H, Gilber D, Wojciechowski R, Alston C, West SK. Nuclear cataract shows significant familial aggregation in an older population after adjustment for possible shared environmental factors. Invest Ophthalmol Vis Sci 2004; 45:2182-6. [PMID: 15223793].

41. Klein AP, Duggal P, Lee KE, O'Neill JA, Klein R, Bailey-Wilson JE, Klein BEK. Polygenic effects and cigarette smoking account for a portion of the familial aggregation of nuclear sclerosis. Am J Epidemiol 2005; 161:707-13. [PMID: 15800262].

42. Saadat M, Farvardin-Jahromi M, Saadat H. Null genotype of glutathione S-transferase M1 is associated with senile cataract susceptibility in non-smoker females. Biochem Biophys Res Commun 2004; 319:1287-91. [PMID: 15194507].

43. Saadat M, Farvardin-Jahromi M. Occupational sunlight exposure, polymorphism of glutathione S-transferase M1, and senile cataract risk. Occup Environ Med 2006; 63:503-4. [PMID: 16551760].

44. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struewing JP, Wolf WA. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011; 4:13-[PMID: 21269473].

45. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, Chen L, Denny JC, Wilke RA, Pathak J, Carrell D, Kho AN, Starren JB. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc 2012; 19:225-34. [PMID: 22319176].

46. Rasmussen LV, Peissig PL, McCarty CA, Starren J. Development of an optical character recognition pipeline for handwritten form fields from an electronic health record. J Am Med Inform Assoc 2012; 19:e90-5. [PMID: 21890871].

47. Zuvich RL, Armstrong LL, Bielinski SJ, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, De Andrade M, Doheny KF, Haines JL, Hayes MG, Jarvik GP, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto ME, McCarty CA, McDavid AN, Mirel DB, Olson LM, Paschall JE, Pugh EW, Rasmussen LV, Rasmussen□Torvik LJ, Turner SD, Wilke RA, Ritchie MD. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. Genet Epidemiol 2011; 35:887-98. [PMID: 22125226].

48. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, De Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet. 2011;Chapter 1:Unit1.19.

49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, Sham PC. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 2007; 81:559-75. [PMID: 17701901].

50. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006; 38:904-9. [PMID: 16862161].

51. Verzi MP, Agarwal P, Brown C, McCulley DJ, Schwarz JJ, Black BL. The transcription factor MEF2C is required for craniofacial development. Dev Cell 2007; 12:645-52. [PMID: 17420000].

52. Sitadevi C, Ramaiah Y, Askari Z. Fructose intolerance associated with congenital cataract. Report of a case. Indian J Pediatr 1968; 35:496-8. [PMID: 5719655].

53. Jin C, Chen J, Meng Q, Carreira V, Tam NNC, Geh E, Karyala S, Ho S-M, Zhou X, Medvedovic M, Xia Y. Deciphering gene expression program of MAP3K1 in mouse eyelid morphogenesis. Dev Biol 2013; 374:96-107. [PMID: 23201579].

54. Pasquale EB. Eph receptors and ephrins in cancer: bidirectional signaling and beyond. Nat Rev Cancer 2010; 10:165-80. [PMID: 20179713].

55. Lin H-J, Huang Y-C, Lin J-M, Wu J-Y, Chen L-A, Lin C-J, Tsui Y-P, Chen C-P, Tsai F-J. Single-nucleotide polymorphisms in chromosome 3p14.1- 3p14.2 are associated with susceptibility of type 2 diabetes with cataract. Mol Vis 2010; 16:1206-14. [PMID: 20664687].

56. Ionides AC, Berry V, Mackay DS, Moore AT, Bhattacharya SS, Shiels A. A locus for autosomal dominant posterior polar cataract on chromosome 1p. Hum Mol Genet 1997; 6:47-51. [PMID: 9002669].

57. Crosslin DR, McDavid A, Weston N, Nelson SC, Zheng X, Hart E, De Andrade M, Kullo IJ, McCarty CA, Doheny KF, Pugh E, Kho A, Hayes MG, Pretel S, Saip A, Ritchie MD, Crawford DC, Crane PK, Newton K, Li R, Mirel DB, Crenshaw A, Larson EB, Carlson CS, Jarvik GP. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Hum Genet [Internet]. 2011 Oct 30; Available from: http://www.ncbi.nlm.nih.gov/pubmed/22037903

58. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, Chai HS, Bastarache L, Zuvich R, Peissig P, Carrell D, Ramirez AH, Pathak J, Wilke RA, Rasmussen L, Wang X, Pacheco JA, Kho AN, Hayes MG, Weston N, Matsumoto M, Kopp PA, Newton KM, Jarvik GP, Li R, Manolio TA, Kullo IJ, Chute CG, Chisholm RL, Larson EB, McCarty CA, Masys DR, Roden DM, De Andrade M. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am J Hum Genet 2011; 89:529-42. [PMID: 21981779].

59. Kullo IJ, Ding K, Shameer K, McCarty CA, Jarvik GP, Denny JC, Ritchie MD, Ye Z, Crosslin DR, Chisholm RL, Manolio TA, Chute CG. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. Am J Hum Genet 2011; 89:131-8. [PMID: 21700265].

60. Turner SD, Berg RL, Linneman JG, Peissig PL, Crawford DC, Denny JC, Roden DM, McCarty CA, Ritchie MD, Wilke RA. Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. PLoS ONE 2011; 6:e19586-[PMID: 21589926].

61. Wilke RA, Berg RL, Linneman JG, Peissig P, Starren J, Ritchie MD, McCarty CA. Quantification of the clinical modifiers impacting high-density lipoprotein cholesterol in the community: Personalized Medicine Research Project. Prev Cardiol 2010; 13:63-8. [PMID: 20377807].

62. Pendergrass SA, Verma SS, Holzinger ER, Moore CB, Wallace J, Dudek SM, Huggins W, Kitchner T, Waudby C, Berg R, McCarty CA, Ritchie MD. Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions

using Biofilter, and gene-environment interactions using the PhenX Toolkit. Pac Symp Biocomput. 2013;147–58.

Articles are provided courtesy of Emory University and the Zhongshan Ophthalmic Center, Sun Yat-sen University, P.R. China. The print version of this article was created on 19 September 2014. This reflects all typographical corrections and errata to the article through that date. Details of any changes may be found in the online version of the article.